Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W517–W521***
DOI: 10.1093/nar/gkh441

LOCnet and LOCtarget: sub-cellular localization for structural genomics targets

Rajesh Nair[1,3,*] and Burkhard Rost[1,2,4]

[1]CUBIC and [2]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA, [3]Department of Physics, Columbia University, 538 West 120th Street, New York, NY 10027, USA and [4]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 Saint Nicholas Avenue, New York, NY 10032, USA

Received February 15, 2004; Revised March 26, 2004; Accepted April 16, 2004

ABSTRACT

LOCtarget is a web server and database that predicts and annotates sub-cellular localization for structural genomics targets; LOCnet is one of the methods used in LOCtarget that can predict sub-cellular localization for all eukaryotic and prokaryotic proteins. Targets are taken from the central registration database for structural genomics, namely, TargetDB. LOCtarget predicts localization through a combination of four different methods: known nuclear localization signals (PredictNLS), homology-based transfer of experimental annotations (LOChom), inference through automatic text analysis of SWISS-PROT keywords (LOCkey) and *de novo* prediction through a system of neural networks (LOCnet). Additionally, we report predictions from SignalP. The final prediction is based on the method with the highest confidence. The web server can be used to predict sub-cellular localization of proteins from their amino acid sequence. The LOCtarget database currently contains localization predictions for all eukaryotic proteins from TargetDB and is updated every week. The server is available at http://www.rostlab.org/services/LOCtarget/.

OVERVIEW

Structural genomics initiatives unravel protein structures

Over 15 structural genomics initiatives currently aim at determining a large number of protein structures in a high-throughput manner. These projects have already deposited almost 700 new protein three-dimensional (3D) structures into the Protein Data Bank (PDB) over the last four years (1). The rate at which structures are experimentally determined for which no low-resolution models are available is over five times higher for structural genomics consortia than it is for the entire PDB (2). One ultimate goal is to experimentally determine at least one representative 3D structure for all sequence–structure families. It is now clear that we need over ten times more structures to realize this concept than was originally anticipated (2; J. Liu and B. Rost, manuscript submitted). Nevertheless, it is currently believed that structural genomics consortia will be able to experimentally determine almost 10 000 new structures before the year 2010. If chosen optimally, these 10 000 would halve the number of residues in known proteins for which we do not have any structural annotations (J. Liu and B. Rost, unpublished). The PDB has created a centralized registration database for target sequences from structural genomics projects worldwide called TargetDB [http://targetdb.pdb.org, (3)]. TargetDB currently contains over 50 000 target sequences. The 3D structure for the majority (>98%) of these sequences is currently unknown and many lack any functional annotations. Functionally annotating structures from structural genomics is currently an important challenge for computational biology (4–6).

Sub-cellular localization one key toward unravelling protein function

Proteins that cooperate toward a common biological function are often located in the same sub-cellular compartment. Aberrant sub-cellular localization of proteins has been observed in the cells of several diseases, such as cancer and Alzheimer's disease. Thus, the sub-cellular localization of a protein is an important coarse-grained aspect of its role. Some of the publicly available predictors for sub-cellular localization are TargetP (7), a neural-network-based method for predicting signal peptides, mitochondrial targeting peptides and chloroplast targeting peptides; neural networks for prediction of the subcellular location of proteins (NNPSL) (8), a neural-network-based localization predictor using amino acid composition;

*To whom correspondence should be addressed. Tel: +1 212 305 4018; Fax: +1 212 305 7932; Email: nair@maple.bioc.columbia.edu

SubLoc (9), a support-vector-based predictor using amino acid composition; and PSORT and PSORT II (10), which are based on identifying sequence motifs responsible for protein sorting, on sequence homology and on NNPSL. We have developed LOCtarget (http://www.rostlab.org/), a database with a web server for the prediction of sub-cellular localization for all sequences in TargetDB. LOCtarget is a comprehensive system for localization prediction based on database annotations for sequence homologues (LOChom), functional information in the form of SWISS-PROT keywords (LOCkey), sequence motifs involved in targeting the nucleus (PredictNLS) and a system of neural networks (LOCnet) for *de novo* prediction. LOCnet was found to be >7% more accurate than the best publicly available system (11). LOCtarget can also be used to predict sub-cellular localization for proteins in any other context. In particular, LOCnet and LOCtarget differ from LOC3D (12) in that they predict for proteins of unknown structures. The LOCtarget database can be useful in complementing other predicted functional information regarding the target sequences in the SPAM database that provides annotations for TargetDB entries [http://spam.sdsc.edu/perl/browser_beta.pl, (13)].

METHODS AND RESULTS

LOCtarget combines the following four different paths to annotate and predict sub-cellular localization (Figure 1). Additionally we report signal peptide predictions using SignalP.

(i) *PredictNLS: identification of nuclear localization signals*. The most accurate way to predict nuclear localization is to identify the nuclear localization signal (NLS). Active transport of proteins into the nucleus is realized by specific molecules such as importins and karyopherins that bind to distinct targeting signals (14). This targeting signal typically contains a short segment of consecutive residues and is commonly referred to as the NLS. PredictNLS (15,16) uses a set of expert-curated experimentally known NLSs to predict nuclear localization. At 100% accuracy this tool identifies about half of all known nuclear proteins.

(ii) *LOCkey: digesting experimental data from SWISS-PROT keywords*. Our second most accurate tool to infer localization uses experimental descriptions of protein function as contained in the controlled vocabulary of SWISS-PROT keywords (17,18). First, we align the target sequence to sequences in SWISS-PROT using pairwise BLAST (19). Second, we extract all SWISS-PROT keywords for all sequence homologues that meet specified thresholds in terms of sequence similarity and the content of these keywords. LOCkey (20) then infers sub-cellular localization through an automated lexical analysis of the extracted SWISS-PROT keywords. In contrast to dictionary-based approaches, LOCkey is fully automated and the rule libraries used to infer localization from keywords are generated dynamically. The method is extremely accurate when any functional information in the form of keywords is known (>82% accuracy using full cross-validation).

(iii) *LOChom: inference through sequence homology*. The next most reliable means of finding out the sub-cellular localization is through homology-transfer. If a protein of experimentally known localization L is significantly sequence similar to a query protein Q, Q and L have identical localization (21,22). We have carried out the most exhaustive study of the sequence conservation of sub-cellular localization to establish the thresholds for annotation transfer based on homology (21). Sequence homologues were first identified using pairwise BLAST and PSI-BLAST. To assign sub-cellular localization three measures of sequence similarity were investigated: pairwise sequence identity, BLAST/PSI-BLAST expectation values (EVAL) and distances from the Sander–Schneider curve that relates alignment length to sequence identity (23,24) (referred to as HSSP-value or HVAL). Of the three measures, the HSSP-value was the most successful in annotating sub-cellular localization. One of the results of our original investigation was a problem-specific refinement for the HSSP-value (20). The use of position-specific scoring matrices in

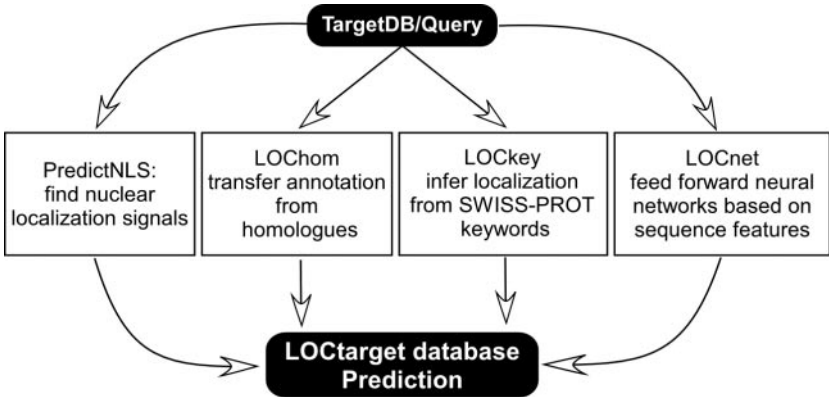

Figure 1. The LOCtarget system. From the query amino acid sequence, the three-state secondary structure and solvent accessible surface residues of the protein are predicted using PROFphd (29; B. Rost, manuscript submitted). LOCtarget uses four different methods to annotate sub-cellular localization. (i) PredictNLS: the amino acid sequence is scanned for nuclear localization signals. (ii) LOChom: the sequence is first aligned through PSI-BLAST profiles to a database with experimental annotations about localization. If any sequence homologues are discovered, sub-cellular localization annotation is transferred from the homologue. (iii) LOCkey: the SWISS-PROT database contains functional information for proteins in the form of keywords. LOCkey infers sub-cellular localization based on keyword entries. The above three programs are based solely on the amino acid sequence of the protein and do not use any structural information. (iv) LOCnet: sub-cellular localization is predicted by a system of neural networks trained on a number of global features such as amino acid composition, predicted secondary structure composition and composition of predicted surface accessible residues. The final localization annotation in the LOCtarget database is taken from the most reliable prediction amongst the four individual methods.

PSI-BLAST also improved the reliability of the homology-transfer. Further improvements in homology-based annotation were obtained through the use of separate 'conservation thresholds' and 'accuracy versus sequence similarity' curves for each of the localization classes. Note that at the level at which we use LOChom, our decisions are for all compartments significantly more accurate than any *de novo* prediction, and even than predictions based on signal or target motifs.

(iv) *LOCnet:* de novo *prediction from sequence*. LOCnet is a system that predicts sub-cellular localization from sequence using neural networks (11). The LOCnet system consists of three layers that sort proteins into one of four classes (extra-cellular, cytoplasmic, nuclear and mitochondrial). Major sources of improvement over publicly available methods originated from using predicted secondary structure [from PROFsec (25,26,27)], improved predictions of solvent accessibility [from PROFacc (28), and evolutionary information from sequence profiles. LOCnet has a module that implicitly predicts generic signal peptides (but not the cleavage sites) and target peptides (11). The final four-state classification accuracy of the system was ~65%. This is nearly 10 percentage points higher than systems using only amino acid composition. We also noted that we had to develop a system tailored specifically to predicting localization for proteins of known 3D structure (11) that is available through LOC3D (12). Although LOCnet performs better for extra-cellular proteins with signal peptides, it can also identify proteins that are secreted through (an) alternative pathway(s) (such as fgf, IL-1), and—in combination with other methods—it can distinguish between proteins with signal peptides that are retained in the endoplasmic reticulum or golgi apparatus and proteins that are actually secreted (11,29). Note, however, that LOCnet is significantly less accurate than TargetP for mitochondrial proteins.

(v) *SignalP: prediction of generic signal peptides*. SignalP (version 2) is a neural-network-based prediction of generic N-terminal signal peptides (7,30). Prediction accuracy for eukaryotic proteins is ~70–80% (11). Note that—due to licensing issues—we do not return the detailed predictions from SignalP; rather we only indicate whether or not SignalP detected a signal peptide.

Best single method determines the final annotation of localization

The final annotation of localization by LOCtarget is taken from the most reliable prediction amongst the four individual methods. Using this four-step approach significantly improves prediction accuracy since different methods are most accurate in different regimes. For example, if an NLS is detected by PredictNLS, the protein has a high probability of being nuclear (our NLS motifs are exclusive to nuclear proteins). If functional information in the form of SWISS-PROT keywords is available, LOCkey can use this information to infer sub-cellular localization at a very high accuracy. In the absence of sufficient functional information, identification of sequence homologues using LOChom proves most accurate. *De novo* predictions using LOCnet are the least accurate means; however, they are applicable when all the other methods fail. In fact, most structural genomics targets could only be predicted by LOCnet (82.7%, Table 1). At the other extreme, the most accurate method (PredictNLS) contributed <3% to the final annotations (Table 1).

Table 1. Annotations of LOCtarget by method

Method	% of proteins
LOCnet	82.7
LOChom	7.6
LOCkey	7.1
PredictNLS	2.6
Total	100.0

Table 2. Annotations by LOCtarget by type of localization

Sub-cellular localization[a]	Eukaryotic sequences	Prokaryotic sequences
Cytoplasm	1142	36 195
Extra-cellular space	764	7594
Nucleus	1816	0
Mitochondria	753	0
Periplasm	68	1287
Chloroplast	53	0
Endoplasmic reticulum	31	0
Golgi apparatus	34	0
Lysosome	12	0
Peroxysome	14	0
Vacuoles	4	0
Signal peptides[b]	525	5745
Total	4691	45 076

[a]Number of target sequences in the LOCtarget database assigned to the given localization.
[b]Method: SignalP 2.0.

Fewer than 10% eukaryotic proteins

The LOCtarget database currently contains sub-cellular localization information for nearly 50 000 targets (Table 2); most of these are from prokaryotes and archae. Note that for the 700 or so proteins for which we have 3D structures, the predictions of LOCtarget and LOC3D may differ, since LOCtarget predictions are based on sequences, not on structures. Of the prokaryotic proteins, ~17% are predicted as extra-cellular (Table 2). Of the eukaryotic proteins in TargetDB, nuclear proteins constitute the single largest group, accounting for ~39% of the sequences. Proteins secreted to the extra-cellular space account for 16% of the proteins, while proteins retained in the cytoplasm account for 24%.

INPUT, OUTPUT AND OPTIONS

Database description

The LOCtarget database has been formatted in an EMBL-like flat-file format. The database can be accessed on the web through a PERL CGI interface. The database can be used in either query mode or browse mode.

(i) *User query*: any object in the database can be queried using a PERL regular expression-like syntax. The query can be a name or a wildcard pattern (the search engine automatically appends the '*' wildcard pattern at the end of the query). If the query field is left blank, the search

displays all objects of the selected type. Three types of object can be queried: TargetDB protein identifiers, types of sub-cellular localization and type of prediction method. For example, querying the 'sub-cellular localization class' object with 'nuclear' displays all proteins in the database that are predicted to have nuclear localization.

(ii) *Browsing the database*: in this mode, database entries are displayed in order of decreasing confidence of prediction.

Web server description

The LOCtarget web server has been implemented using a PERL CGI interface. Currently sequences can be submitted only in FASTA format. However, we anticipate that the server will accept any standard sequence format in the very near future (FASTA, PIR, MSF, SWISS-PROT) or a list of protein identifier codes from SWISS-PROT, TrEMBL or PDB. We will also enable uploading sequences from the user's local machine. Users have the option of receiving plain text (ASCII) output or HTML-formatted results that can be displayed in any web browser. Results are returned as email attachments.

Format and fields

Each protein can have up to four localization predictions associated with it, one from each method. The database uses four fields to represent predictions from each method:

(i) *Method*: the type of prediction method used.

(ii) *Loci*: predicted sub-cellular localization from this method. The predicted sub-cellular localization can be one of nine classes (Table 2).

(iii) *Confidence*: confidence score assigned by the prediction method. This is a number between 0 and 100. Larger confidence scores mark more accurate predictions.

(iv) *Details*: any reasons, if available, for the particular localization class inferred by the method. For example, for a LOCkey prediction, this field would give details of the keywords responsible for this localization prediction.

FUTURE

We are currently extending our system to also predict localization for prokaryotic and archaeal targets. Next, we will incorporate our annotations of sub-cellular localization prediction into the SGTDB/SPAM database (http://spam.sdsc.edu/perl/browser_beta.pl), which provides annotations for TargetDB. Further extensions will include entirely sequenced organisms and improved prediction methods (R. Nair and B. Rost, in preparation).

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu and Megan Restuccia (Columbia) for computer assistance and to Kaz Wrzeszczynski (Columbia) for valuable discussions. Thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego University), John Westbrook (Rutgers) and their crews for maintaining excellent databases and to all experimentalists who enabled this tool by making their data publicly available. The work of R.N. and B.R. was supported by the grant DBI-0131168 from the National Science Foundation (NSF) and the grant R01-LM07329-01 from the National Library of Medicine (NLM).

REFERENCES

1. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al*. (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
2. Liu,J., Hegyi,H., Acton,T.B., Montelione,G.T. and Rost,B. (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins*, (in press).
3. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
4. Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3341–3344.
5. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2003) From protein structure to biochemical function?. *J. Struct. Funct. Genomics*, **4**, 167–177.
6. Goldsmith-Fischman,S. and Honig,B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci.*, **12**, 1813–1821.
7. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
8. Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2235.
9. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
10. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *TIBS*, **24**, 34–36.
11. Nair,R. and Rost,B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, **53**, 917–930.
12. Nair,R. and Rost,B. (2003) LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Res.*, **31**, 3337–3340.
13. Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al*. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
14. Tinland,B., Koukolikova-Nicola,Z., Hall,M.N. and Hohn,B. (1992) The T-DNA-linked VirD2 protein contains two distinct functional nuclear localization signals. *Proc. Natl Acad. Sci., USA*, **89**, 7442–7446.
15. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localisation signals. *EMBO Rep.*, **1**, 411–415.
16. Nair,R., Carter,P. and Rost,B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res.*, **31**, 397–399.
17. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
18. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
19. Altschul,S., Madden,T., Shaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
20. Nair,R. and Rost,B. (2002) Inferring sub-cellular localisation through automated lexical analysis. *Bioinformatics*, **18**, S78–S86.
21. Nair,R. and Rost,B. (2002) Sequence conserved for sub-cellular localization. *Protein Sci.*, **11**, 2836–2847.
22. Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *TICB*, **8**, 169–170.
23. Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

24. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*., **12**, 85–94.
25. Rost,B. (2001) Protein secondary structure prediction continues to rise. *J. Struct. Biol*., **134**, 204–218.
26. Rost,B. and Liu,J. (2003) The PredictProtein server. *Nucleic Acids Res*., **31**, 3300–3304.
27. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol*., **266**, 525–539.
28. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein server. *Nucleic Acids Res*., **32**, W321–W326.
29. Wrzeszczynski,K.O. and Rost,B. (2004) Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes. *CMLS*., in press.
30. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*., **10**, 1–6.